SO-Sequence Ontology Update:
Results of Sequence Ontology Workshop on Non-canonical features

18-19 Sept 2006
Sanger Institute, Hinxton, UK

BRC4 Dec 2006
Jessie Kissinger
Jkissing@uga.edu

Participants

- Michael Ashburner
- Suzi Lewis
- Karen Eilbeck
- Ariane Toussaint
- Anne Summers
- Jessie Kissinger

miSO
sequence ontology term browser

Search for terms in SO.

Type a word to search for in SO. The autocomplete function returns a list of matches.
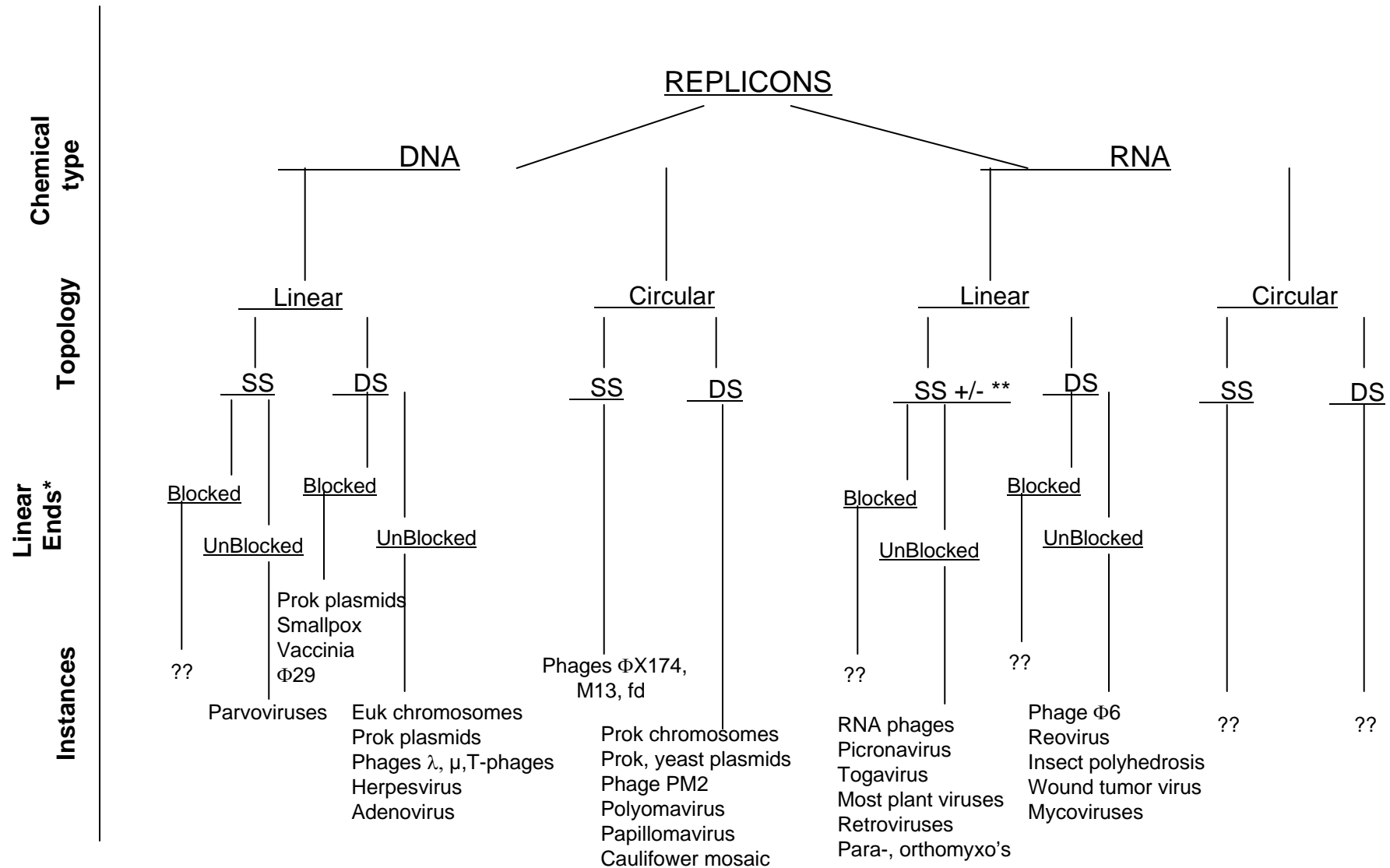
chromosome

locate term

Browse SO

miSO provides a graphical means to browse the terms in the Sequence Ontology. The view is organised via a parent/child approach. The terms are all linked to the SO term tables web page which displays more details such as the relationship types, the definition, the synonyms and the cross references.

Sequence_Ontology
⊞ consequences_of_mutation
⊞ chromosome_variation

chromosome	SO:0000340

Definition: Structural unit composed of long DNA molecule.

DBxref:external reference [http:////biotech.icmb.utexas.edu/search/dict-search.mhtml]

Synonyms:

Aspect: located_sequence_feature

Parent relationships: chromosome **is_a** region	Child relationships: chromosome_region **part_of** chromosome

chromosome_arm	SO:0000105

Definition: A region of the chromosome between the centromere and the telomere. Human chromosomes have two arms, the p arm (short) and the q arm (long) which are separated from each other by the centromere.

DBxref:external reference [http:////www.exactsciences.com/cic/glossary/_index.htm]

Synonyms:

Aspect: located_sequence_feature

Parent relationships: chromosome_arm **is_a** chromosome_region	Child relationships:

http://www.sequenceontology.org/miSO/header.html

A HIERARCHY OF REPLICONS BASED ON CHEMISTRY AND TOPOLOGY

Chemical type

REPLICONS

DNA — RNA

Topology

Linear — Circular — Linear — Circular

Linear Ends*

SS — DS (under DNA Linear)
SS — DS (under DNA Circular)
SS +/- ** — DS (under RNA Linear)
SS — DS (under RNA Circular)

Blocked / UnBlocked (under various)

Instances

DNA Linear SS:
- Blocked: ??
- UnBlocked: Parvoviruses

DNA Linear DS:
- Blocked: Prok plasmids, Smallpox, Vaccinia, Φ29
- UnBlocked: Euk chromosomes, Prok plasmids, Phages λ, μ, T-phages, Herpesvirus, Adenovirus

DNA Circular SS:
- Phages ΦX174, M13, fd

DNA Circular DS:
- Prok chromosomes, Prok, yeast plasmids, Phage PM2, Polyomavirus, Papillomavirus, Caulifower mosaic

RNA Linear SS +/- **:
- Blocked: ??
- UnBlocked: RNA phages, Picronavirus, Togavirus, Most plant viruses, Retroviruses, Para-, orthomyxo's

RNA Linear DS:
- Blocked: ??
- UnBlocked: Phage Φ6, Reovirus, Insect polyhedrosis, Wound tumor virus, Mycoviruses

RNA Circular SS: ??

RNA Circular DS: ??

*Ends of linear replicons can either be free (unblocked) 5'phosphates and 3' hydroxyls or blocked in short hairpins (in SS) or covalently crosslinked (DS).

** SS RNA replicons can either be directly translatable mRNA's (+) or complementary to mRNA (-), needing to be copied before used as mRNA.

AO Summers, UGA, Nov 2006

Basic and Mosaic MGE's

Agents of HGT, MGE (pure prototypes)	Plasmids (RP4, RK2, R100, ColE1)	Phages (lambda)	Transposons (Tn5, Tn10, Tn501)
Mosaic MGE's	P1 A temperate (lysogenic) phage that is a plasmid	Mu A temperate (lysogenic) phage that that replicates by transposition	Conjugative Transposon A transposon that conjugates to other cells, e.g. Tn916
	Genomic Islands have conjugative, phage-derived, and transposon-derived components all located together on the main chromsome. Pathogenicity islands (PAIs) have virulence genes. Also called ICE, integrative conjugative element (20KB - >100s KB)		

MGE's have 3 defining processes

		1. Self-Replication	2. Transfer	3. Site-specific Recombination
Genes	Trans-acting: Enzymes and NA binding proteins	Replicases, NA polymerases, rolling circle replication, phage 'host-takeover' functions	DNA "pump", a Type IV secretion system (T4SS). Phage packaging proteins	Transposases, integrases, excisases, resolvases
	Cis-acting: NA interaction or cleavage sites	oriV	oriT, phage packaging sites, cos site	Gene cassettes, Att sites, Tnp IR's, Res sites

AO Summers, UGA, Nov 2006

MGE's also carry "baggage" loci: functions that are valuable but not intrinsic to being a gene transfer agent.

Examples of "baggage" genes		1. Pathogenicity	2. Metabolism	3. Resistance
Genes	Trans-acting: Enzymes, transport proteins and NA binding proteins	Toxins , invasion proteins, colonization factors	Catabolic pathways for xenobiotic compounds. Nitrogen fixation, photosynthesis	Antibiotic resistance, toxic metal resistance.
	Cis-acting: NA interaction sites	Relevant operators, promoters, etc.	Relevant operators, promoters, etc	Relevant promoters, operators, etc

Some "baggage" loci reside within transposons that are carried by plasmids or phages.

Ontology Needs & Benefits for Mobile Genetic Elements

- **Missing <u>gene</u> ontologies for intrinsic MGE functions, e.g.:**
  - DNA pump (variant of general secretion systems)
  - Recombinases (transposases, resolvases, integrases)
  - Phage structural/packaging genes
    - Tail fibers
    - Capsid proteins (self-assemble into phage protein coat)
- **Missing <u>sequence</u> ontologies for intrinsic MGE functions, e.g.:**
  - Transfer origins, *oriT* and unique replication origins, *oriV*
  - *Res* sites for transposition
  - *Att* sites for integration
  - Phage packaging sites
- **Benefits of improved MGE ontology for the community:**
  - Literature searching
  - Find novel relationships, insights
  - Facilitate expert annotation - A.I. makes it easier and faster to incorporate their expertise

New Chromosome Definintion

Plasmids, not just reagents any more

Some Mobile Elements Added

How to get new terms

- Make a few ppt slides
  - Define the biological problem, pictures help
  - Define the existing terms and their shortcomings
  - Propose new terms and their definitions
- Send the above to Karen Eilbeck, kareneilbeck@mac.com
- Get OBO-Edit from Sourceforge.net download current SO from: http://www.sequenceontology.org/so.shtml

Our current use of SO

- Our usage of available terms is sparse and disparate betweenBRCs
- BRCs use ~14 terms in our GFF files:

  (contig, mRNA, tRNA, rRNA, gene, CDS, gene, five_prime_utr, three_prime_utr, region, exon, repeat_region, mature_peptide, ribosome_entry_site)

- GenBank has 50+ terms
- SO has 200+ terms (SOFA has ~100)
- Should we decide on standards and depth? Just because we can annotate it, should we?
- What do our communities need?
- View Sample GFF -SOFA term variation